KATSAUS

Biografiasampo yhdistää ja rikastaa suomalaiset elämäkerrat linkitettynä datana semanttisessa webissä

Eero Hyvönen
eero.hyvonen@aalto.fi
https://orcid.org/0000-0003-1695-5840

Petri Leskinen
petri.leskinen@aalto.fi
https://orcid.org/0000-0003-2327-6942

Minna Tamper
minna.tamper@aalto.fi
https://orcid.org/0000-0002-3301-1705

Heikki Rantala
heikki.rantala@aalto.fi
https://orcid.org/0000-0002-4716-6564

Esko Ikkala
esko.ikkala@aalto.fi
https://orcid.org/0000-0002-9571-7260

Jouni Tuominen
jouni.tuominen@aalto.fi
https://orcid.org/0000-0003-4789-5676

Kirsi Keravuori
kirsi.keravuori@finlit.fi
https://orcid.org/0000-0003-3288-7606

Aalto-yliopisto, Helsingin yliopisto (HELDIG-keskus) ja Suomalaisen Kirjallisuuden Seura
https://seco.cs.aalto.fi/projects/biografiasampo/

Informaatiotutkimuksen tavoitteena on kehittää uusia tapoja tuottaa, organisoida ja käyttää tietoa sekä yksilöiden että organisaatioiden näkökulmasta. Tässä katsauksessa esitellään kulttuurihistoriallisen tiedon tuottajia ja käyttäjiä palvelevan ns. Sampo-mallin sovellus Biografiasampo kansalaisille, digitaalisten ihmistieteiden tutkijoille ja uusien sovellusten kehittäjille. Biografiasammon kunnianhimoisena tavoitteena on käynnistää uusi aikakausi elämäkertakokoelmien julkaisemisessa ja käyttämisessä verkossa semanttisen webin teknologioita ja linkitetyn avoimen datan julkaisuperiaatteita hyödyntäen. Innovaationa on luoda kieliteknologian, tekoälyn ja semanttisen webin teknologioiden avulla elämäkertojen teksteistä ja niihin eri lähteissä liittyvistä tietokannoista tietämysverkko (knowledge graph) osana kansallista tietoinfrastruktuuria. Sovelluksen ydinaineistona ovat Kansallisbiografia ja muut Suomalaisen Kirjallisuuden Seuran toimittamat ja julkaisemat pienoiselämäkerrat, yhteensä 13 900 elämäntarinaa, joita on kirjoittanut 980 suomalaista tutkijaa maamme suurimmaksi sanotussa historiantutkimuksen hankkeessa. Elämäkerroista louhittua dataa on rikastettu automaattisen loogisen päättelyn avulla ja linkittämällä sitä 16 muuhun tietolähteeseen. Tietämysverkko on julkaistu linkitetyn avoimen datan Linked Data Finland -palvelussa. Datapalvelun avulla on toteutettu seitsemästä sovellusnäkymästä koostuva älykäs, avoin ja maksuton verkkopalvelu biografiasampo.fi, jolla on ollut noin 50 000 käyttäjää. Sekä järjestelmän elämäkerrat että niistä louhittu data ovat avoimesti käytettävissä datapalveluna Linked Data Finland -alustalla.

Asiasanat: elämäkerrat, semanttinen web, prosopografia, yhditetty avoin tieto, portaalit (tietotekniikka)

Pysyvä osoite: https://doi.org/10.23978/inf.107948

Sampoja verkossa

Sampo-mallin (Hyvönen, 2021) ja siihen perustuvan Biografiasammon ytimessä on linkitetyn datan idea yleisöllisestä tiedon julkaisemisesta (Heath & Bizer, 2011) ja käyttämisestä, jossa kaikki voivat voittaa: tiedon julkaisijat voivat rikastaa sisältöjään ”ilmaiseksi” toisten julkaisijoiden dataa linkittämällä ja uutta tietoa päättelemällä, ja loppukäyttäjille voidaan tarjota aiempaa runsaampia tietosisältöjä aiempaa älykkäämpien käyttöliittyminen ja työkalujen kautta. Sampo-mallin mukaisten portaalien (Sampo-portaalit, 2021) käyttöliittymien toteutuksessa (Ikkala et al., 2021) keskeinen idea on ollut fasettihaun (Tunkelang, 2009, Tzitzikas et al., 2017) yhdistäminen data-analyyttisiin työkaluihin. Siinä hakukohteista suodatetaan ensin esiin kiinnostuksen kohteena oleva joukko kohteita tekemällä valintoja hierarkkisista faseteista, jotka sammoissa perustuvat ontologioihin (Studer & Staab, 2009). Tämän jälkeen tulosjoukkoa voidaan tutkia tarkemmin kohdistamalla siihen mm. tilastollisia analyysejä, visualisointeja ja verkostoanalyysiä. Avointa yhteistä linkitetyn datan infrastruktuuria (Hyvönen, 2020b) yhä uudelleen hyödyntämällä ja asteittain kehittämällä uusien sampojen

kehittäminen on saatu kustannustehokkaaksi, kun pyörää ei tarvitse keksiä joka kerta uudestaan. Lisäksi W3C:n standardeja käyttämällä voidaan rakentaa siltaa impivaarasta kansainvälisen semanttisen webin infrastruktuureihin.

Tässä katsauksessa esitellään esimerkkinä Sampo-mallin sovelluksista järjestelmä *Biografiasampo – suomalaiset elämäkerrat semanttisessa webissä*, joka koostuu linkitetyn avoimen datan palvelusta ja sen varaan kehitetystä semanttisen webin portaalista. Tarkastelemme aluksi järjestelmää linkitetyn tiedon tuotannon näkökulmasta ja sitten datapalvelun ja portaalin käyttöä elämäkerroista kiinnostuneiden kansalaisten sekä digitaalisten ihmistieteiden tutkijoiden kannalta.

Kansalliset elämäkertakokoelmat

Suomalaisen Kirjallisuuden Seuran (SKS) tiedekustantamo julkaisee Suomen historian keskeisten henkilöiden pienoiselämäkertoja kirjoina ja verkossa. SKS on tuottanut yhteistyössä Suomen Historiallisen Seuran kanssa vuonna 1997 julkaistun Kansallisbiografia-verkkopalvelun (Kansallisbiografia, 2021), johon kuuluu 6500 elämäkertaa. Ne julkaistiin 10-osaisena, 9500-sivuisena suurteossarjana vuosina 2003–2008. SKS:n aineistoihin kuuluvat myös erilliset tietokannat mm. Suomen papistosta, Venäjän sotavoimissa palvelleista kenraaleista ja amiraaleista sekä talouselämän vaikuttajista. Aineistoja laajennetaan uusilla elämäkerroilla yhdessä muiden tieteellisten seurojen kanssa (ks. taulukko 1). Aineistot ovat olleet luettavissa verkossa osin maksullisen palvelun kautta.

Taulukko 1. SKS:n henkilöhistorialliset tietokannat, jotka muodostavat Biografiasammon ytimen.

	Elämäkertakokoelma	**Koko**
1	Kansallisbiografia	6512
2	Kenraalit ja amiraalit 1809–1917	481
3	Suomen papisto 1800–1920	1953
4	Talouselämän vaikuttajat	2238
5	Turun hiippakunnan paimenmuisto 1554–1721	2716
	Yhteensä elämäkertoja	**13900**

Vastaavia, vielä laajempia kirja- ja verkkopalveluhankkeita on muissakin maissa, esimerkkeinä Britannian Oxford Dictionary of National Biography

(ODNB, 2021), USA:n American National Biography (ANB, 2021), Saksan Neue Deutsche Biographie (NDB, 2021), Alankomaiden Biography Portal of the Netherlands (BPN, 2021) ja BiographyNet (2021) ja Ruotsin Svenskt Biografiskt Lexikon (SBL, 2021). Alan ehkä tunnetuin kokoelma ODNB sisältää n. 63 000 elämäkertaa ja Biography Portal of the Netherlands lähes 150 000. Elämäkertojen pituus eri kokoelmissa vaihtelee matrikkelimaisista yhteenvedoista seikkaperäisiin artikkeleihin.

Henkilöhistoriallinen tieto on tärkeä komponentti historian tutkimuksessa, ja elämäkerrallisten tietojen saatavuus kätevästi yhdestä auktoritatiivisesta kokoelmasta helpottaa tutkijoiden työtä. Kansallisten biografiahankkeiden yhtenä pontimena on perinteisesti ollut myös kansallisvaltioiden identiteetin lujittaminen tarkastelemalla historiaa omien suurmiesten ja -naisten toiminnan kautta. Biografia kiinnostaa tutkijoiden ohella myös suurta yleisöä, ja elämäkerrat ovat usein kirjakauppojen myyntitilastojen kärkisijoilla. Akateemisen historiantutkimuksen uusia painotuksia seuraten elämäkertakokoelmien näkökulma on laajentunut suurmiehistä ja -naisista kohti arjen historiaa, aiemmin tuntemattomia vaikuttajia ja uusien ryhmien kuten eri vähemmistöjen edustajia. Elämäkerrallista tietoa julkaistaan verkossa biografiakokoelmien ohella runsaasti muuallakin, kuten Wikipediassa, organisaatioiden ja ammattikuntien matrikkeleissa, kansainvälisissä auktoriteettitietokannoissa, sukututkimuksen piirissä sekä museoiden ja arkistojen kokoelmissa.

Biografiasammon visio

SKS:n julkaisemat suomalaiset elämäkerrat ja vastaavat kansainväliset kokoelmat ovat olleet lähes poikkeuksetta saatavilla vain tekstimuodossa ihmisen luettavaksi joko painettuna tai sähköisesti verkossa, mutta ei semanttisena eli tietokoneen ”ymmärtämänä” datana (Heath & Bizer, 2011; Hyvönen, 2018). Lisäksi tietokannoissa oleva metatieto rajoittuu yleensä biografisiin perustietoihin ja suuri osa tiedosta on olemassa vain rakenteettomana artikkelitekstinä, jota tietokone ei ”ymmärrä”. Elämäkertatietojen analysoimiseen ja tutkimiseen ei juuri ole tarjolla työkaluja tai palveluita yksinkertaisia, yhteen tietokantaan kohdistuvia ajan, paikan, sukupuolen ja ammatin sekä niiden yhdistelmien mukaisia hakuja lukuun ottamatta. Kehittyneemmät data-analyyttiset työkalut olisivat kuitenkin tarpeen henkilöhistoriallisten aineistojen tutkimusta varten digitaalisissa ihmistieteissä (Gardiner & Musto, 2015). SKS:n elämäkerroilla on ollut verkossa vuosittain n. 280 000 lukijaa, joten voi arvioida, että kysyntää on myös kehittyneille palveluille.

Biografiasammon (Hyvönen et al., 2019; Biografiasampo, 2021) uutuusarvo perustuu laskennallisten menetelmien uudenlaiseen hyödyntämiseen sekä biografioiden julkaisussa että niiden tarjoamisessa loppukäyttäjille helppokäyttöisinä työkaluina. Ideana on 1) parantaa aineistojen haku- ja selailuominaisuuksia ja mahdollistaa erilaisten ihmisryhmien joustava muodostaminen prosopografista tutkimusta varten, 2) rikastaa elämäkerrallisia aineistoja toistensa avulla sekä parantaa ja monipuolistaa näin käyttäjän lukukokemusta, 3) tarjota lukijalle välineet yksittäisen elämäkerran parempaan hahmottamiseen, visualisoimiseen ja biografiseen tutkimiseen, 4) mahdollistaa ihmisryhmien ominaisuuksien prosopografinen ja tilastollinen tutkimus ja vertailu, 5) muodostaa uutta tietämystä tekoälyn avulla (knowledge discovery) sekä 6) helpottaa mahdollisuuksia tutkia ja vertailla elämäkerroissa käytettyä kieltä. Biografiasampo demonstroi paradigman muutosta, jossa elämäkertojen julkaisemisessa on siirrytty ensin painetuista kirjoista haku- ja selailutoiminnoilla varustettuun verkkojulkaisemiseen. Seuraava askel on aineistojen julkaiseminen datana data-analyyttisiin välineisiin integroituina palveluina. Nyt ollaan siirtymässä tekoälyperustaisiin järjestelmiin, joissa kone ei ole vain passiivinen työkalu vaan voi aktiivisesti auttaa myös tutkimuskysymysten etsinnässä, ratkaisemisessa ja niiden selittämisessä käyttäjälle (Hyvönen, 2020). Douglas Adamsin klassikkoromaanissa *Linnunradan käsikirja liftareille* (Hitchhikers Guide to the Galaxy) tietokoneelta haluttiin vastaus kysymykseen elämästä, maailmankaikkeudesta ja kaikesta muusta sellaisesta. 7,5 miljoonan vuoden laskennan jälkeen saatu vastaus ”42” voi olla oikein, mutta jäi epäselväksi, mikä oikeastaan oli kysymys, ja tutkija kuulisi mielellään myös perustelun vastaukselle.

Biografiasampo koostuu kahdesta osasta, 1) semanttisesta portaalista ja 2) linkitetyn avoimen datan palvelusta Linked Data Finland -alustalla (Hyvönen et al., 2014; LDF, 2021). Biografiasampo.fi-portaali tarjoaa käyttäjälleen älykkäät haku- ja selailutoiminnot, joihin on saumattomasti integroitu joukko data-analyyttisiä työkaluja ja visualisointeja henkilöhistoriallisten aineistojen tutkimista ja analysointia varten verkostoina, tilastoina, erilaisina graafeina ja kartoilla. Portaalin käyttö ei edellytä erityistä tietoteknistä osaamista. Biografiasammon datapalvelun avoimet rajapinnat ja SPARQL-palvelupiste tarjoavat helppokäyttöisen mahdollisuuden uusien data-analyysien toteuttamiseen digitaalisten ihmistieteiden tutkijoille, joilla on jonkin verran kokemusta semanttisen webin SPARQL-kyselykielestä ja/tai ohjelmoinnista esimerkiksi Jupyter- ja Google Colab -dokumenttien avulla Python-kielellä.

Tekstistä dataan

Biografiasammon elämäkerrat on talletettu SKS:n henkilöhistoriallisiin tietokantoihin. Kustakin henkilöstä on olemassa rakenteisessa muodossa biografiset perustiedot kuten henkilön nimi, ammatti sekä syntymä- ja kuolinvuosi ja -paikka. SKS:n perinteisessä verkkopalvelussa (Kansallisbiografia, 2021) pienoiselämäkertoja voidaan hakea henkilön perustiedoilla hakulomakkeella ja sitten valita hakutulosjoukosta yksi biografia kerrallaan luettavaksi ja tutkittavaksi (close reading, lähiluku). Biografiateksteihin on lisätty linkkejä toisiin biografioihin tarpeen mukaan, mikä mahdollistaa jossain määrin myös aineiston eksploratiivista selailua (Marchionini, 2006). Aineistojen automatisoitu data-analyysi ja visualisointi on vaikeaa (distant reading, kaukoluku) (Moretti, 2013) eikä henkilöitä voida tutkia prosopografisesti ryhminä (Verboven, 2007) laskennallisesti, esimerkiksi ammattikunnittain tai aikakauden perusteella.

Yksittäisen elämäkerran teksti sisältää ensin lyhyen ingressin ja vapaasti kirjoitetun kuvauksen henkilön elämästä, joita seuraa puoliformaaliin tapaan matrikkelityylillä kirjoitettu yhteenveto henkilön perhesuhteista, urasta, teoksista, saavutuksista yms. Esimerkiksi arkkitehti Eliel Saarisen (1873–1950) yhteenveto-osa on esitetty alla lyhennettynä (…):

> Gottlieb Eliel Saarinen S 20.8.1873 Rantasalmi, K 1.7.1950 Bloomfield Hills, Michigan, Yhdysvallat. V rovasti Juho Saarinen ja Selma Maria Broms. P1 1898 - 1902 (ero) Mathilda Tony Charlotta Gyldén (sittemmin Gesellius) S 1877, K 1921, P1 V agronomi Axel Gyldén ja Antonia Sofia Hausen; P2 1904 - kuvanveistäjä Minna Carolina Louise (Loja) Gesellius S 1879, K 1968, P2 V liikemies Herman Otto Gesellius ja Emilie Struckmann. Lapset: Eva-Lisa (Pipsan) S 1905, K 1979, sisustussuunnittelija, P arkkitehti Jons Robert Ferdinand Swanson; Eero S 1910, K 1961, arkkitehti.
>
> URA. Käynyt kaksi luokkaa Viipurin suomalaista klassillista lyseota 1883 - 1887, kolme luokkaa Viipurin Alkeiskoulua 1887 - 1890; ylioppilas Tampereen reaalilyseosta 1893; arkkitehti Suomen Polyteknillisestä opistosta 1897; yliopiston piirustuskoulu 1894 - 1897; Sjöströmin stipendiaatti Saksassa, Ruotsissa 1898 - 1899; opintomatkoja Euroopan eri maihin, Yhdysvaltoihin.
>
> Arkkitehtitoimisto Gesellius, Lindgren & Saarinen, perustajajäsen, osakas 1896 - 1907; Arkkitehtitoimisto Eliel Saarinen, johtaja 1907 - 1923; arkkitehti Evanstonissa (Illinois) 1923 -1924, Ann Arborissa (Michigan) 1924 - 1937; Arkkitehtitoimisto Eliel & Eero Saarinen 1937 - 1941; partneri, Saarinen-Swanson-Saarinen and Associates 1941 - 1947, Saarinen, Saarinen and Associates 1947 - .

Arkkitehtuurin vieraileva professori Michiganin yliopistossa (Ann Arbor) 1924 - 1925; Cranbrookin taideakatemian suunnittelija, opettaja 1925 - , johtaja 1932 - .

Yhdysvaltain kansalainen 1945.

Jäsenyydet: Suomen Taideakatemia; Ruotsin taideakatemia; Pietarin taideakatemia.

Kunnianosoitukset: Suomen Leijonan suurr.; Suomen Valkoisen Ruusun K I. Professorin arvo 1919. Kunniatohtori: Helsinki 1932, Teknillinen korkeakoulu 1934, Karlsruhe 1933, Michiganin yliopisto 1933, Harvardin yliopisto 1940, Cambridgen yliopisto 1940, Draken yliopisto 1948, Des Moinesin yliopisto 1948. Kultamitali: Architectural League of New York 1933, ...

TEOKSET. Katso M. Hausen, K. Mikkola, A.-L. Amberg, T. Valto, Eliel Saarinen : Suomen aika. 1990.

Arkkitehtitoimisto Gesellius, Lindgren, Saarinen: Tallbergin talo. 1896 - 1898, Luotsikatu 1, Helsinki; Pariisin maailmannäyttelyn 1900 paviljonki. 1898 - 1900, Pariisi; Vakuutusyhtiö Pohjolan talo. 1899 - 1901, Mikonkatu 3, Helsinki; Pohjoismaiden Osakepankki. 1903 - 1904, Unioninkatu 32, Helsinki (purettu 193..4); Suomen Kansallismuseo. 1902 - 1911, Helsinki; Helsingin Työväenyhdistyksen talo. 1904, ...

TUOTANTO. The city, its growth, its decay, its future. New York 1943; Search for Form : a fundamental approach to art. New York 1947.

LÄHTEET JA KIRJALLISUUS. A.-L. Amberg, Saarisen sisustustaide 1896 - 1923. 1984; Ars : Suomen taide 4. 1989: R. Wäre, Arkkitehtuuri vuosisadan vaihteessa; A. Christ-Janer, Eliel Saarinen : Finnish-American Architect and Educator. Chicago, London 1979 (Revised Edition); ...

ELIEL SAARISEN MUKAAN NIMETTY. Eliel Saarisen tie 1952, Elielinaukio 1996, Helsinki; Eliel Saarisen puisto, Joensuu; postimerkki 1973; ravintola Eliel, Helsingin rautatieasema.

AUKTORITEETTITUNNISTEET. VIAF: 39471157

Biografiasammossa tällaisista tiivistelmistä ja elämäkertojen vapaasta tekstiosuudesta louhittiin ja linkitettiin nimettyjä entiteettejä (kuten paikkojen ja henkilöiden nimiä), asiasanoja ja tapahtumia (event). Tapahtumat kuvattiin CIDOC CRM -ontologiamallista (Doerr, 2003) ja standardista (CRM, 2021) laajennetulla Bio CRM -tietomallilla (Tuominen et al., 2017). Siinä tapahtumia luonnehtivat niihin eri rooleissa osallistuneet toimijat (henkilöt,

ryhmät ja organisaatiot), paikat ja ajat. Ideana oli kuvata henkilöiden elämä niiden tapahtumien sarjana, joihin hän on osallistunut syntymästä kuolemaan ja sen jälkeenkin postuumisti. Eri tietolähteistä eri muodoissa saatavaa dataa harmonisointiin yhteentoimivaksi tapahtumatasolla, mikä mahdollisti aineistojen sisällöllisen rikastamisen toistensa avulla. Esimerkiksi Ateneumin taidemuseosta saatava perinteinen Dublin Core -tyyppinen (DC, 2021) metatietokuvaus maalauksesta voidaan esittää taiteellisena luomistapahtumana ja yhdistää se sitten taiteilijan elämäkerran erilaisten tapahtumien jatkumoon. Perhesuhteista rakennettiin oma risteilevien sukupuiden verkosto, jota rikastettiin päättelemällä epäsuoria sukulaissuhteita, kuten täti tai pikkuserkku (Leskinen & Hyvönen, 2019). Elämäkertojen vapaille teksteille tehtiin kieliteknologinen analyysi ja muunnettiin ne lingvistiseksi tietämysverkoksi, jonka varaan voidaan tehdä kielellisiä analyysejä, joita on esitelty tarkemmin lähteessä (Tamper et al., 2018). Tekstit myös linkitettiin niissä mainittuihin ontologisiin entiteetteihin, erityisesti henkilöihin ja paikkoihin mm. verkostoanalyysejä varten.

Elämäkerrallisesta ja lingvistisestä datasta muodostettiin linkitetyn datan palvelu Linked Data Finland -alustalle, jonka rajapintojen varaan toteutettiin sovelluksena semanttinen portaali Biografisampo osoitteessa `https://biografiasampo.fi`.

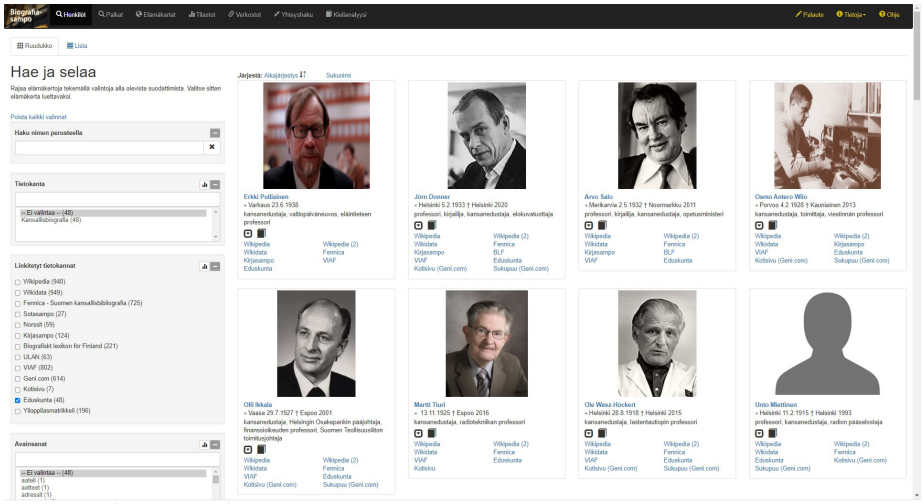
Esimerkkejä Biografiasammon käytöstä

Seuraavassa havainnollistetaan Biografiasammon avaamia uusia mahdollisuuksia käytännön esimerkkien avulla. Esittelemme sekä portaalin käyttöliittymän tarjoamia valmiita haku-, selailu- ja data-analyysin välineitä että datapalvelun käyttöä digitaalisten ihmistieteiden tutkijoille ja sovellusten kehittäjille.

1) Hakuominaisuuksien parantaminen: henkilöt, ryhmät ja paikat

Aineistoista louhittu tietämysverkko (knowledge graph) (Noy et al., 2019) mahdollistaa elämäkertojen ja ihmisryhmien haun joustavasti ns. fasettihaun (Tunkelang, 2009; Tzitzikas et al., 2017) avulla, joka perustuu S. R. Ranganathanin jo 1930-luvulla ideoimaan fasettiluokituksen teoriaan kirjastotieteessä. Siinä elämäkertoja voi suodattaa perinteisen nimihaun lisäksi tekemällä valintoja ontologisten luokitusten avulla. Niitä ovat valittu SKS:n tietokanta, linkitetty muu tietolähde (kuten Wikidata tai eduskunnan kansanedustajamatrikkeli), avainsanat, ajanjakso, biografian kirjoittaja, toimiala,

yritys tai yhteisö, arvo, ammatti tai toiminta ja syntymäpaikka. Esimerkiksi Getty-tutkimuskeskuksen historiallisten henkilöiden ULAN-rekisteriin (ULAN, 2021) päässeet suomalaiset tai Helsingin yliopiston ylioppilasmatrikkelissa 1853–1899 dokumentoidut henkilöt löytyvät yhdellä klikkauksella ”linkitetyt tietokannat” -fasetista. Professorit ja muiden ammattien edustajat taas suodattuvat esiin vastaavalla valinnalla ammattien ja arvojen fasetista. Fasettivalintojen kohdalla olevat, automaattisesti päivittyvät lukumäärät kertovat, kuinka monta elämäkertaa löytyy, jos tekee kyseisen valinnan. Näin lukija ei koskaan päädy tilanteeseen, jossa haun tulos on tyhjä joukko (jos hakija tätä ei halua), mikä on perinteisiin hakumenetelmiin liittyvä harmillinen ominaisuus. Kuvassa 1 käyttäjä on valinnut ammatti tai arvo -fasetista kategorian ”professori” ja sitten linkitettyjen tietokantojen fasetista ”eduskunta”, jolloin tuloksena on 48 professoria tai professorin arvonimen saanutta henkilöä, jotka ovat toimineet myös kansanedustajana.



Kuva 1. Kansanedustajien elämäkertojen haku fasettihaulla.

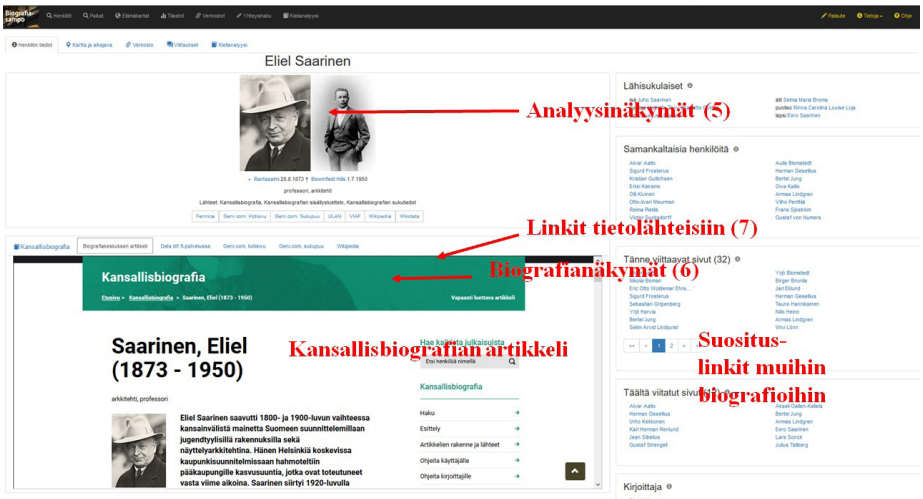
Henkilöiden ja ihmisryhmien haun lisäksi Biografiasampo tarjoaa koko maapallon kattavan karttaperustaisen hakunäkymän, jossa elämäkerroista louhitut kymmenet tuhannet tapahtumat on projisoitu kartalle n. 2600 eri paikkaan. Näkyviin saa myös historiallisia karttoja ja paikkoja esimerkiksi luovutetun Karjalan alueelta. Karttapalvelu perustuu Suomalaisten historiallisten paikkojen ja karttojen ontologiapalveluun Hipla.fi (Hyvönen et al., 2016).

2) Elämäkertojen ja lukukokemuksen rikastaminen

Biografiasampo rikastaa käyttäjän lukukokemusta eri tavoin linkittämällä dataa automaattisesti eri lähteistä. Kone myös päättelee algoritmien avulla uutta tietoa analysoimalla kokonaisuutta. Henkilöistä tarjotaan elämäkerrallisia kuvauksia SKS:n aineistojen lisäksi myös muista tietolähteistä, kuten Wikipediasta, Kirjasammosta, sukututkijoiden aineistoista (Geni.com) ja Getty-tutkimuskeskuksen ULAN-rekistereistä (ks. taulukko 2). Ajatuksena on rekonstruoida henkilöiden elämätarinat mahdollisimman monen eri tietolähteen tietoja yhdistämällä ja luomalla jokaiselle henkilöille automaattisesti ”kotisivut”. Järjestelmän tekoäly ikään kuin lukee elämäkerrat ja tutkii niihin liittyvät tietokannat sekä tarjoaa tästä jäsennetyn yhteenvedon lukijalle edelleen tutkittavaksi. Ilman tietokoneen apua näin laajojen ja heterogeenisten aineistojen lukeminen ja jäsentäminen olisi erittäin työlästä tai mahdotonta.

Taulukko 2. Biografiasammon elämäkertoihin linkitetyt muut tietolähteet, joita voi käyttää fasettihaun kriteerinä.

	Tietolähde/palvelu	**Linkit**	**Kuvaus/osoite**
1	Wikipedia	6507	`http://fi.wikipedia.org`
2	Wikidata	6972	`http://www.wikidata.org`
3	Fennica	4007	Kansalliskirjaston kansallisbibliografia, linkitetyn datan palvelu
4	Sotasampo	287	`http://sotasampo.fi`
5	Vanhat Norssit	221	`http://www.norssit.fi/semweb`
6	Kirjasampo	715	`http://kirjasampo.fi`
7	Biografiskt lexikon för Finland	1305	Ruotsinkieliset biografiat (SLS), `http://www.blf.fi`
8	ULAN-ontologia	229	`http://www.getty.edu/research/tools/vocabularies/ulan`
9	VIAF	3121	Virtual International Authority FIles, `http://viaf.org`
10	Geni.com	5321	Kansainvälinen sukututkijoiden sukupuu-palvelu
11	Kotisivu	43	Henkilökohtaiset kotisivut
12	Eduskunta	614	Kansanedustajat 1917–2018
13	Ylioppilasmatrikkelit 1640–1852 ja 1853–1899	2046	Helsingin yliopiston tiedekuntien opiskelijat ja jäsenet
	Yhteensä	**31388**	



Kuva 2. Eliel Saarisen kotisivu Biografiasammossa.

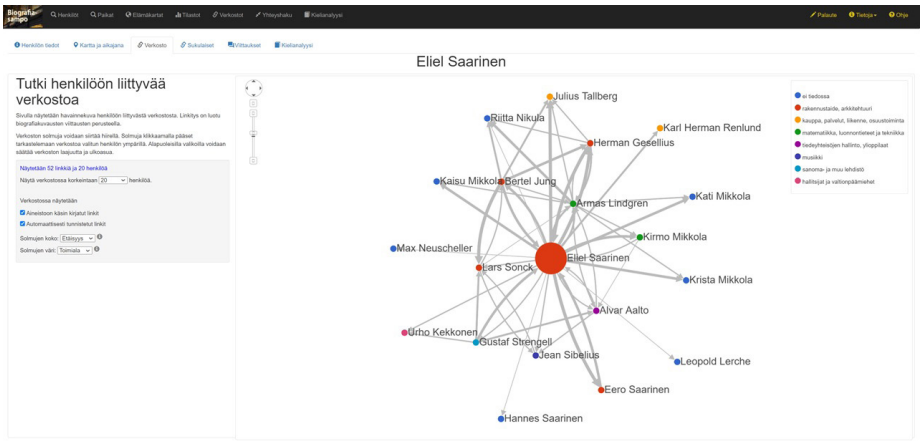
Kuvassa 2 näkyy esimerkkinä Eliel Saariselle automaattisesti muodostettu kotisivu. Ylärivissä on avattavissa viisi erilaista välilehteä, joiden kautta voi tarkastella Saarisen elämää eri tavoilla: tietosivuna (valittuna kuvassa), elämäntapahtumina kartalla ja aikajanalla, Saarisen egosentrisenä, elämäkertojen välisiin viittauksiin perustuvana verkostona, viittausanalyysina muihin elämäkertoihin sekä Kansallisbiografian pienoiselämäkerran kielianalyysina.

Saarisen eri lähteistä löydettyjen kuvien ja perustietojen alla on linkkejä erilaisiin tietolähteisiin, joissa Saarisesta on tietoa, tässä tapauksessa seitsemään eri verkkopalveluun Suomessa ja ulkomailla. Hänestä on Biografiasammossa kuusi erilaista biografista kuvausta, jotka ovat luettavissa eri välilehdiltä: SKS:n julkaisema teksti ja alkuperäinen artikkelisivu, Geni.com-palvelun kotisivu ja sukupuunäkymä, Wikipedian artikkeli sekä Ylioppilasmatrikkelin 1640–1899 tiedot. Näistä alkuperäinen Kansallisbiografian artikkelisivu on kuvassa avattu luettavaksi. Kuvan oikeassa reunassa on vielä tietokoneen luomia erilaisia suosituslinkkejä Eliel Saariseen eri tavoin liittyviin muihin sivuihin, esimerkiksi toisiin sisällöltään samankaltaisiin elämäntarinoihin, jotka on tunnistettu automaattisesti niiden sanaston ja sen merkityksen perusteella.

3) Biografinen tutkimus

Käyttäjälle tarjotaan elämäkertatekstien ohella joustava ja helppokäyttöinen data-analyyttinen välineistö (kuvan 1 analyysinäkymät), jolla voi tutkia ja visualisoida yksittäisen henkilön elämäntarinaa. Esimerkiksi kuvassa 3 on Eliel Saarisen ns. egosentrinen verkko, josta näkyvät hänen yhteytensä artikkeleissa

tehtyjen viittausten perusteella muihin elämäkertojen päähenkilöihin, kuten Hvitträskin toimiston arkkitehtikollegoihin Herman Geselliukseen ja Armas Lindgreniin. Valitsemalla toisen välilehden visualisoituvat hänen kansainvälisen elämänsä tapahtumat tyypeittäin jaoteltuna kartalla ja aikajanalla. Biografiasammon kieliteknologinen tekoäly louhii tapahtumat automaattisesti elämäkerrasta ja muista tiedoista (Leskinen et al., 2018; Tamper et al., 2018).



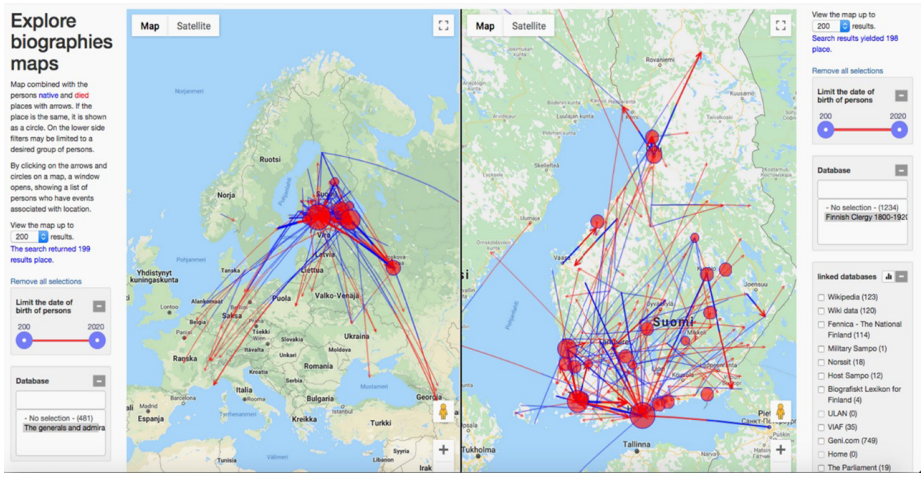
Kuva 3. Eliel Saarisen egosentrinen verkosto, joka on tässä rajoitettu 20 henkilöön. Henkilöiden toimialat on merkitty eri väreillä.

4) Prosopografinen tutkimus ja tilastot

Samalla tavalla Biografiasammossa voi tutkia ja verrata toisiinsa henkilöryhmien toimintaa prosopografian keinoin. Esimerkiksi Norssissa vuosina 1867–1992 opiskelleiden oppilaiden myöhemmin muodostamia verkostoja voi luoda ja visualisoida Biografiasampoon linkitettyjen SKS:n aineistojen sisäisten viittausten perusteella. Esiin nousee silloin mm. erillinen piiri filosofi Eino Kailan ympärillä sekä luokkatoverien Matti Klinge, Pentti Saarikoski ja Anto Leikola muodostama verkosto. Näin Biografiasampo itseasiassa rikastaa automaattisesti vuonna 2017 julkaistun Vanhat Norssit semanttisessa webissä -palvelun (Hyvönen et al., 2017) toimintoja pelkästään palveluiden henkilöiden välisen siltauksen avulla.

Kuvassa 4 käyttäjä vertaa toisiinsa kahden ihmisryhmän elämänlankoja, Venäjän sotavoimissa 1809–1917 palvelleita suomalaisia amiraaleja ja kenraaleja (vasemmalla) ja suomalaista papistoa vuosina 1800–1920 (oikealla). Ryhmät on muodostettu kahdella rinnakkaisella fasettihaulla Biografiasammon elämäkarttojen vertailunäkymässä, jossa elämä kuvataan sinipunaisena nuolena syntymäpaikasta (sininen pää) kuolinpaikkaan (punainen pää).

Yhdellä vilkaisulla selviää, että upseerit liikkuivat pappeja kansainvälisemmin ja kohti etelää kuten eläkeläiset nykyään. Yhtä kaarta kartalla klikkaamalla pääsee käsiksi kaareen liittyviin elämäkertoihin tarkempaa tutkimusta varten. Esimerkiksi vasemmalla näkyvä poikkeava kaari Oulusta Länsi-Siperiaan osoittautuu Siperiaan maanmittaustöiden johtajaksi nimitetyn kenraali Gustav Adolf Silverhjelmin (1799–1864) elämänkaareksi.



Kuva 4. Venäjän sotavoimissa Suomen suuruhtinaskunnan aikana palvelleiden suomalaisten kenraalien ja amiraalien (vasemmalla) ja papiston (oikealla) elämänkaarien prosopografinen vertailu Biografiasammossa.

Biografiasampoon sisältyy lisäksi koko joukko tilastollisia näkymiä ihmisryhmien analysoimista varten. Pylväsdiagramminäkymän avulla selviää mm. faseteilla valitun henkilöryhmän ikäjakautuma, lasten ja puolisoiden lukumäärät sekä elinajat ja iät. Piirakkadiagrammien avulla taas saa kuvan henkilöryhmän jakautumisesta sukupuolen, ammatin, toimialan, ammatin tai arvon sekä yrityksen tai yhteisön mukaan. Myös kuvan 4 kaltaiset eri ryhmien väliset tilastolliset vertailut ovat mahdollisia.

5) Tietämyksen muodostus tekoälyllä

Semanttisen webin ja tekoälyn suuria lupauksia on uuden tietämyksen automaattinen löytäminen (knowledge discovery) linkitetyistä aineistoista, jopa serendipiteetti (Pease et al., 2013). Linkitetyn datan kokonaisuus on usein jo sellaisenaan enemmän kuin osiensa summa. Tarkasti määritelty tietämysverkon looginen semantiikka mahdollistaa lisäksi uusien tietojen välisten yhteyksien päättelyn ja lisäämisen verkkoon. Biografiasammossa tätä on

hyödynnetty mm. fasettihaun ontologioissa, jossa vaikkapa ”Saksassa” syntyneitä etsittäessä löytyy Friedenaussa syntynyt, toisen maailmansodan aikana Mikkelin päämajassa toiminut saksalainen yhdysesikunnan kenraali Waldemar Erfurth, koska Friedenau on osa Berliiniä, joka puolestaan on osa Saksaa. Järjestelmään sisältyvässä erillisessä Yhteyshaku-sovelluksessa (Hyvönen & Rantala, 2021) puolestaan tavoitteena on päätellä ja löytää henkilöiden ja paikkojen välisiä yhteyksiä ja muodostaa niille suomenkieliset selitykset. Esimerkiksi fasettivalinnoilla ”Italia” ja ”taidemaalari” löytyvät suomalaisten taidemaalarien erilaiset semanttiset yhteydet Italiaan, kuten että Elin Danielson-Gambogi vastaanotti Firenzen kaupungin taidepalkinnon vuonna 1899 ja että ”Robert Wilhelm Ekman on luonut vuonna 1844 taideteoksen 'Maisema Subiacosta', joka kuvaa paikkaa Italia”. Jälkimmäisessä tapauksessa yhteys on muodostunut Biografiasampoon yhdistettyjen Ateneumin kokoelmatietojen kautta. Ateneumin lisäksi yhteyshaussa hyödynnetään lisäksi kokoelmadataa kansallisbibliografia Fennicasta (2021), Kirjasammosta (2021), historiallisia tapahtumia kuvaavasta HISTO-ontologiasta (2021) sekä J. V. Snellmanin kootuista teoksista (SKT, 2021), joista luotiin linkitetyn datan versio.

6) Tekstien kielellinen analyysi

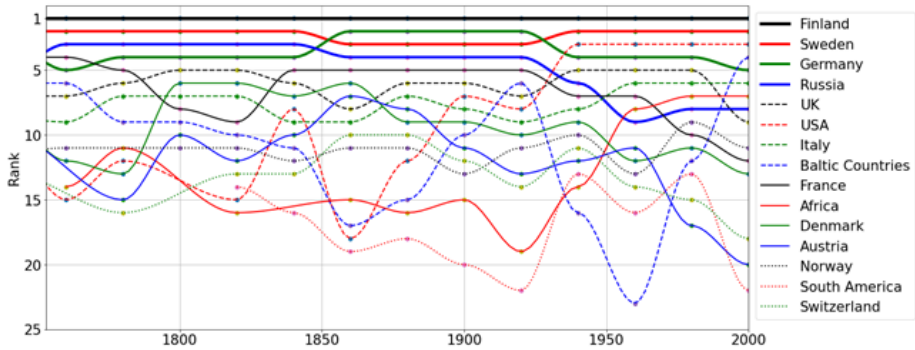
Biografioita voi tutkia Biografiasammossa paitsi henkilöiden, tapahtumien ja paikkojen myös teksteissä käytetyn kielen kautta (Tamper et al., 2018). Aineistoista on muodostettu erillinen yli 100 miljoona yhteyttä sisältävä kielellinen tietämysverkko, jossa tekstien sanamuodot on perusmuotoistettu ja analysoitu sekä lauseet jäsennetty. Tämän datan avulla järjestelmään on toteutettu mm. työkalu, jolla voidaan tutkia ja verrata eri henkilöryhmien biografioissa käytettyjä sanoja. Esimerkiksi itsenäisyyden ajan nais- ja mieskansanedustajien elämäkertoja voi verrata vastaavan tyyppisellä vertailunäkymällä kuin kuvassa 4. Näkymä luo tilaston sanojen käytöstä ja kertoo, että esimerkiksi sanat ”perhe” ja ”lapsi” ovat erittäin yleisiä naisia kuvaavissa elämäkerroissa, mutta eivät juuri esiinny miesten kuvauksissa.

Järjestelmässä on myös palvelu, josta selviävät ne lauseet, kontekstit, joiden kautta biografioiden välinen viittausverkosto ja siihen perustuvat egosentriset verkostot eri henkilöille on muodostettu. Näistä lauseista saa selityksensä esimerkiksi verkostoissa näkyvä yllättävä yhteys keihäänheittäjä Tapio Rautavaaran ja runoilija, akateemikko Aale Tynnin välillä. Yhteys syntyy siitä, että molemmat voittivat olympiakultaa Lontoossa, Rautavaara keihäänheitossa ja Tynni lyriikassa, joka oli vielä tuolloin olympialaji.

Avoimen datapalvelun hyödyntäminen

Biografiasampo perustuu semanttisen webin linkitetyn datan (Hyvönen, 2018) standardeihin (W3C, 2021) ja julkaisumalliin (Heath & Bizer, 2011). Siinä data julkaistaan aktiivisena datapalveluna, jota sovellukset hyödyntävät SPARQL-kyselykielen ja muiden rajapintojen kautta. Datapalvelun selkeä erottaminen sovelluksista mahdollistaa saman datan hyödyntämisen eri tavoin, mitä jo Biografiasampoon sisältyvät seitsemän eri sovellusnäkymää demonstroivat. Sovelluksissa toteutettujen haku-, selailu- ja data-analyyttisten työkalujen ohella datapalvelua voi käyttää myös suoraan rajapintojen kautta, kuten SPARQL-kieltä tukevalla YASGUI-järjestelmällä (Rietveld & Hoekstra, 2017) tai Python-skriptien ja visualisointien avulla Googlen Colab- (2021) ja Jupyter-järjestelmien (Jupyter, 2021) kautta. Näin toteutettavien analyysien rajana on vain tutkijan mielikuvitus ja tietysti käytettävissä olevan datan ominaisuudet, kattavuus ja laatu.

Esimerkiksi kuvan 5 visualisointi on tehty Google Colabin Python-ympäristössä Biografiasammon avoimen datapalvelun SPARQL-rajapinnan varaan (Tamper et al., 2021). Siinä on laskettu Kansallisbiografian elämäkerroissa eri vuosiin liittyen mainitut maat ja valittu ja järjestetty niistä mittauspisteissä aina 15 eniten mainittua. Graafista saa mielikuvaa eri maiden suhteellisesta merkityksestä Suomessa eri aikoina.



Kuva 5. Eri valtioiden mainintojen määriin Kansallisbiografiassa perustuva historiografinen visualisointi, joka on toteutettu Google Colab -järjestelmällä ja Biografiasammon linkitetyn avoimen datan palvelulla. Sen mukaan esimerkiksi eri aikoina eläneiden henkilöiden elämäkerroissa tehdyt viittaukset Ranskaan ja Itävaltaan ovat viime aikoina vähentyneet, maininnat Baltian maista taas ovat lisääntyneet huomattavasti.

Digitaalista lähdekritiikkiä tarvitaan

Biografiasammossa olevat SKS:n aineistot perustuvat asiantuntijoiden laatimiin kirjoituksiin, mutta tietojen rakenteistamisen, koostamisen, yhdistelyn, rikastamisen ja uuden tiedon muodostamisen on tehnyt etupäässä tietokone. Koska aineistot ovat laajoja, ei kokonaisuuden virheettömyyttä voida tarkistaa käsin kuin testimielessä sieltä täältä. Linkityksistä löytyy siksi enemmän virheitä kuin jos ne olisi tehty käsityönä. Tilastollisten ja verkostoanalyysien johtopäätösten kanssa on syytä olla tarkkana ja pitää mielessä, mihin dataan ja laskentamenetelmään ne perustuvat. Tämä on tyypillistä digitaalisissa ihmistieteissä, jossa käsitellään usein niin laajoja aineistoja, ettei systemaattinen ihmistyö ole mahdollista aineistoja muodostettaessa ja analysoitaessa. Laskennallisten tekniikoiden lisäarvo on kuitenkin erityisen suuri tällaista suurdataa (big data) käsiteltäessä ja puutteellinenkin tulos monasti parempi kuin ei mitään tulosta. Biografiasampoa pyritään jatkokehittämään digitaalisten ihmistieteiden tutkijoiden työkaluksi lisäämällä tarkempaa tietoa niin aineistojen lähtökohdista, synnystä ja rajoituksista kuin yksittäisten työkalujen toiminnasta.

Elämäkertoja datana tutkittaessa on huomattava, että data perustuu elämäkertakokoelmaan, joka vain epäsuorasti heijastelee reaalimaailman ilmiöitä. Esimerkiksi toimituskunnan valinnat siitä, kenestä elämäkertoja kirjoitetaan, vaikuttavat ratkaisevasti dataan. Datan analyysi on siksi luonteeltaan historiografista ja avaa kiinnostavia uudenlaisia näkymiä myös elämäkertakokoelmien luomisprosessiin (Tamper et al., 2021). Elämäkertojen biografinen ja prosopografinen data-analyysi nostaa kuitenkin esiin myös taustalla olevaan historiaan liittyviä kiinnostavia ilmiöitä, joiden todellisuutta voidaan ryhtyä tutkimaan tarkemmin perinteisen historiantutkimuksen menetelmin. Vastaavanlaista historiografista tutkimusta on aiemmin tehty Iso-Britannian ODNB-biografioista ja Irlannin kansallisbiografioista (Warren et al., 2016; Warren, 2018; Bhreathnach et al., 2019).

Biografiasammon kaltaisen järjestelmän käyttäjältä edellytetään uudenlaista datalukutaitoa (data literacy) ja erityisen tarkkaa lähdekriittistä ymmärrystä siitä, millaista dataa tietoaineisto ja sovellus oikeastaan sisältävät, missä määrin tieto on epätäsmällistä tai puutteellista ja millaisia oletuksia järjestelmässä käytettävät ontologiat ja menetelmät kenties sisältävät (Koltay, 2015; Mäkelä et al., 2020; Université du Luxembourg, 2021). Esimerkiksi faseteissa käytetyssä paikkojen ontologiassa luovutetun Karjalan paikat eivät löydy Suomen alta, vaikka niiden avulla kuvatut tapahtumat yleensä liittyvä aikaan, jollain alue vielä oli osa Suomea.

Huomattava on myös esimerkiksi se, että Biografiasammon verkosto-analyysien perustana ovat SKS:n julkaisemien elämäkertojen väliset linkit, jotka aineistojen kirjoittajat ja toimittajat ovat luoneet. Siksi linkki ei tarkoita sitä, että yhdistetyt henkilöt olisivat välttämättä edes tienneet toisistaan. Esimerkiksi Blanka Namurilaisen (Ruotsin kuningatar 1318–1363), Ruotsin ja Norjan kuninkaan Maunu Eerikinpojan ranskalaisen puolison verkosto kuvaa paitsi hänen lähipiiriään myös hänen jälkimainettaan. Oman aikansa ruotsalaisen ylimystön lisäksi kuningatar Blankan egosentrisen verkoston keskeisiä linkkejä ovat viisisataa vuotta myöhemmin eläneet Zachris Topelius ja Albert Edelfelt. Topeliuksen Lukemisia lapsille ja Edelfeltin maalaus tekivät Blankasta yhden tunnetuimmista ruotsalaisista kuningattarista.

Biografiasammon kaltainen järjestelmä ei korvaa perinteistä tekstien lähilukua mutta helpottaa big data -aineistojen hakua, selailua ja analyysiä sekä auttaa löytämään aineistosta kiinnostavia ilmiöitä tarkempaa tutkimusta varten.

Biografiasammossa sovellettu ja edelleen kehitetty linkitetyn datan tuotannon, julkaisemisen ja käytön Sampo-malli (Hyvönen, 2021), on kehitetty Aalto-yliopiston tietotekniikan laitoksen ja Helsingin yliopiston Digitaalisten ihmistieteiden keskuksen HELDIGin Semanttisen laskennan tutkimusryhmässä (SeCo). Työ on osa laajempaa kieliteknologian ja semanttisen laskennan tutkimusta, jonka tavoitteena on kansallisen linkitetyn avoimen datan tietoinfrastruktuurin (LODI4DH, 2021) luominen ja soveltaminen digitaalisten ihmistieteiden tutkimukseen (Hyvönen, 2020b). Biografiasampo on jäsen Sampo-järjestelmien sarjassa (Sampo-portaalit, 2021), jotka on toteutettu asteittain kehittyvän infrastruktuuriin varaan ja jotka testaavat ja demonstroivat Sampo-mallin käyttökelpoisuutta. Aiemmin julkaistuja sampoja ovat mm. Kulttuurisampo (2021) (monialaiset kulttuurikokoelmat, julkaistu 2008), Kirjasampo (2021) (yleisten kirjastojen kertomakirjallisuus ym., julkaistu 2011), Sotasampo (2021) (Kansallisarkiston, Puolustusvoimien ym. toisen maailmansodan aineistot, 742 000 käyttäjää, julkaistu 2015), Nimisampo (2021) (mm. Kotuksen Nimiarkisto ja Maanmittauslaitoksen paikannimirekisteri, julkaistu 2018), Sotasurmasampo 1914–1922 (2021) (sisällissodan, heimosotien ja 1. maailmansodan uhrit ja taistelut, julkaistu 2019), Mapping Manuscript Migrations (MMM, 2021) (keskiaikaiset käsikirjoitukset Oxfordin Bodleian-kirjastosta, USA:n Schoenberg-instituutista ja Ranskan French Institute for the Research and History of Texts (IRHT) -tutkimuslaitoksesta, julkaistu 2020), Akatemiasampo (2021) (Turun akatemian ja Helsingin yliopiston ylioppilasmatrikkelit 1640–1899, julkaistu 2021) sekä Löytösampo (2021) (Museoviraston arkeologiset, metallinetsintään liittyvät kokoelmat, julkaistu 2021). Valmistumassa ovat myös

Kirjesampo (2021) (historiallista kirjedataa Oxfordin yliopistosta, Hollannista ja Saksasta), Lakisampo (2021) (Suomen lainsäädäntö ja oikeuskäytäntö oikeusministeriön Finlex-tietokannasta) ja Parlamenttisampo (2021) (eduskunnan täysistunnot ja parlamentaarikkojen verkostot). Sammoista käytetyin on nykyisin yleisten kirjastojen ylläpitämä Kirjasampo, jolla sivuilla on ollut kaksi miljoonaa vuosittaista vierailijaa, ja toiseksi suosituin Sotasampo, jonka sivuilla on käynyt yli 740 000 käyttäjää. Biografiasampoa on käyttänyt 50 000 vierasta.

Lisätietoa Biografiasammosta

Biografiasampo-hankkeen kotisivut: `https://seco.cs.aalto.fi/projects/biografiasampo/`

Videoita Biografiasampoon ja linkitetyn avoimen datan infrastruktuuriin liittyen:

- Semantic Web and AI for Digital Humanities: `https://vimeo.com/470313703`
- BiographySampo - AI Reading Biographies for the Semantic Web: `https://vimeo.com/328419960`
- Building a National Level Linked Open Data Infrastructure for Digital Humanities in Finland: `https://vimeo.com/460086143`

Kirjoittajat

Eero Hyvönen on professori tietotekniikan laitoksella Aalto-yliopistossa ja Helsingin yliopiston digitaalisten ihmistieteiden keskuksen HELDIG johtaja. Petri Leskinen, Minna Tamper, Esko Ikkala ja Heikki Rantala ovat tohtorikoulutettavia Aalto-yliopistossa. Jouni Tuominen toimii Helsingin yliopiston HELDIG- ja HSSH-keskuksissa sekä Aalto-yliopistossa staff scientist -tutkijana ja Kirsi Keravuori on SKS:n tiedekustantamon johtaja.

Kiitokset

Biogafiasampo kehitettiin osana Tekesin ja yrityskonsortion rahoittamaa Severi-projektia (2021) vuosina 2016–2018. Työ on jatkunut tämän jälkeen mm. EU:n rahoittamassa kansainvälisessä InTaVia-projektissa (2021) ja

Suomen Akatemian Semparl-projektissa (2021). Laskentaympäristön Biografiasammolle on tarjonnut CSC – Tieteen tietotekniikan keskus Oy.

Kirjallisuutta

Akatemiasampo (14.6.2021). Akatemiasampo-projekti: https://seco.cs.aalto.fi/projects/yo-matrikkelit/

ANB (14.6.2021). American National Biography. Verkkopalvelu osoitteessa: http://www.anb.org/

Biografiasampo (14.6.2021). Hankkeen kotisivu aineistoineen osoitteessa: http://seco.cs.aalto.fi/projects/biografiasampo

BiographyNet (14.6.2021). Palvelu osoitteessa: http://www.biographynet.nl/

Bhreathnach, Ú., Burke, C., Fhinn, J. M., Cleircín, G. Ó., & Raghallaigh, B. Ó. (2019). A Quantitative Analysis of Biographical Data from Ainm, the Irish-language Biographical Database. Teoksessa *Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019)*. CEUR Workshop Proceedings.

BPN (14.6.2021). Biography Portal of the Netherlands. Palvelu osoitteessa: http://www.biografischportaal.nl/en

CRM (14.6.2021). CIDOC-CRM-standardin kotisivut: http://www.cidoc-crm.org/

DC (14.6.2021). Dublin Core Metadata Inititive -kotisivut: https://dublincore.org/

Doerr, M. (2003). The CIDOC CRM—an Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine, 24*(3), 75–92.

Fennica (14.6.2021). Fennica Linked Data: https://www.kiwi.fi/display/Datacatalog/Fennica+Linked+Data

Gardiner, E., & Musto, R. G. (2015). *The Digital Humanities: A Primer for Students and Scholars*. Cambridge University Press. https://doi.org/10.1017/CBO9781139003865

Colab (14.6.2021). Google Colab -palvelu: https://colab.research.google.com/notebooks/intro.ipyn

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology, 1*(1), 1–136. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

HISTO (14.6.2021). Historiaontologia HISTO:n kotisivu: https://seco.cs.aalto.fi/ontologies/histo/

Hyvönen, E., Tuominen, J., Alonen, M., & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. Teoksessa *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers* (pp. 226–230). Springer. https://doi.org/10.1007/978-3-319-11955-7_24

Hyvönen, E. (2018). *Semanttinen web. Linkitetyn avoimen datan käsikirja*. Gaudeamus.

Hyvönen, E. (2020a). Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web, 11*(1), 187–193. https://doi.org/10.3233/sw-190386

Hyvönen, E. (2020b). Linked Open Data Infrastructure for Digital Humanities in Finland. DHN 2020 Digital Humanities in the Nordic Countries. Teoksessa *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference* (pp. 254–259). CEUR Workshop Proceedings, Vol. 2612. http://ceur-ws.org/Vol-2612/

Hyvönen, E. (2021). Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Submitted for peer-review. Pre-print: https://seco.cs.aalto.fi/publications/2021/hyvonen-sampo-model-2021.pdf

Hyvönen, E. (2021). Sammon taontaa semanttisessa webissä. *Tekniikan Waiheita, 39*(2), 87–105. https://doi.org/10.33355/tw.102864

Hyvönen, E., Ikkala, E. & Tuominen, J. (2016). Linked Data Brokering Service for Historical Places and Maps. Teoksessa *Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe)* (pp. 39–52). CEUR Workshop Proceedings, Vol 1608. http://ceur-ws.org/Vol-1608/

Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., & Sirola, L. (2017). Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. Teoksessa J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann (eds.), *Language, Data, and Knowledge* (pp. 113–119). LDK 2017. Lecture Notes in Computer Science, vol 10318. Springer. https://link.springer.com/chapter/10.1007/978-3-319-59888-8_9

Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., & Keravuori, K. (2019). BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. Teoksessa P. Hitzler et al. (eds.), *The Semantic Web* (pp. 574–589). ESWC 2019. Lecture Notes in Computer Science, vol 11503. Springer. https://doi.org/10.1007/978-3-030-21348-0_37

Hyvönen, E., & Rantala, H. (2021). Knowledge-based Relation Discovery in Cultural Heritage Knowledge Graphs. Digital Scholarship in the Humanities (DSH). Accepted. Pre-print: https://seco.cs.aalto.fi/publications/2021/hyvonen-rantala-dsh-2021.pdf

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2021). Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web*, Pre-press. https://doi.org/10.3233/SW-210428

InTaVia (13.9.2021). Projektin 2020–2022 kotisivut: https://seco.cs.aalto.fi/projects/intavia/

Jupyter (14.6.2021). Jupyter Notebooks: https://jupyter.org

Kansallisbiografia (14.6.2021). Käytettävissä vuodesta 1997 alkaen osoitteessa: https://kansallisbiografia.fi/

Kirjasampo (14.6.2021). Käytettävissä osoitteessa: https://www.kirjasampo.fi/

Kulttuurisampo (14.6.2021). Kulttuurisampo-projektin kotisivu: https://seco.cs.aalto.fi/applications/kulttuurisampo/

Klinge, M. (toim.) (2008). *Suomen kansallisbiografia 1–10*. Suomalaisen Kirjallisuuden Seura, 2003–2008.

Koltay, T. (2015). Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science, 49*(1), 3–14. https://doi.org/10.1177/0961000615616450

Lakisampo (14.6.2021). Lakisampo-projektin kotisivu: https://seco.cs.aalto.fi/projects/lakisampo/

LDF (4.6.2021) Linked Data Finland -palvelualusta osoitteessa: https://ldf.fi. Biografiasammon datapalvelu, sen dokumentointi ja SPARQL-palvelupiste löytyvät sivulta: https://www.ldf.fi/dataset/nbf

Leskinen, P., & Hyvönen, E. (2019). Extracting Genealogical Networks of Linked Data from Biographical Texts. Teoksessa P. Hitzler (eds.), *The Semantic Web: ESWC 2019 Satellite Events* (pp. 121–125). Springer. https://doi.org/10.1007/978-3-030-32327-1_24

Leskinen, P., Hyvönen, E., & Jouni Tuominen, J. (2017). Analyzing and Visualizing Prosopographical Linked Data Based on Biographies. Teoksessa A. Fokkens et al. (eds.), *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)* (pp. 39–44). CEUR Workshop Proceedings, Vol. 2119. http://ceur-ws.org/Vol-2119/

Leskinen, P., Miyakita, G., Koho, M., & Hyvönen, E. (2018). Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint. Teoksessa V. Ivanova et al. (eds.), *Proceedings of the Fourth International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 17th International Semantic Web Conference (ISWC 2018)* (pp. 53–63). CEUR Workshop Proceedings, Vol. 2187. http://ceur-ws.org/Vol-2187/

LODI4DH (14.6.2021). Linked Open Data Infrastructure for Digital Humanities -hankkeen kotisivu: http://seco.cs.aalto.fi/projects/lodi4dh

Löytösampo (14.6.2021). Löytösampo-projektin kotisivu: https://seco.cs.aalto.fi/projects/sualt/

Marchionini, G. (2006). Exploratory Search: from Finding to Understanding. *Communications of the ACM, 49*(4), 41–46. https://doi.org/10.1145/1121949.1121979

MMM (14.6.2021) Mapping Manuscript Micrations -projektin kotisivu: https://seco.cs.aalto.fi/projects/mmm/

Moretti, F. (2013). *Distant Reading*. Verso Books.

Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M, . . . Nevalainen, T. (2020). Wrangling with Non-standard Data. Teoksessa S. Reinsone et al. (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference* (pp. 81–96). CEUR Workshop Proceedings, Vol. 2612. http://ceur-ws.org/Vol-2612/

NDB (14.6.2021). Neu Deutsche Biografie. Palvelu osoitteessa: https://www.ndb.badw-muenchen.de/

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: lessons and challenges. *Communications of the ACM, 62*(8), 36–43. https://doi.org/10.1145/3331166

Nimisampo (14.6.2021). Nimisampo-projektin kotisivu: https://seco.cs.aalto.fi/projects/nimisampo/

ODNB (14.6.2021). Oxford Dictionary of National Biography -palvelu: http://global.oup.com/oxforddnb/info

Parlamenttisampo (14.6.2021). Parlamenttisampo-hankkeen kotisivu: https://seco.cs.aalto.fi/projects/semparl/

Pease, A., Colton, S., Ramezani, R., Charnley, J., & Reed, K. (2013). A discussion on serendipity on creative systems. Teoksessa *Proceedings of the Fourth International Conference on Computational Creativity, ICCC 2013* (pp. 64–71). University of Sydney. http://www.computational-creativity.net/iccc2013/download/iccc2013-pease-et-al.pdf

Rietveld, L., & Hoekstra, R. (2017). The YASGUI Family of SPARQL Clients. *Semantic Web, 8*(3), 373–383. https://doi.org/10.3233/SW-150197

RLL (14.6.2021). LetterSampo-projektin kotisivu: https://seco.cs.aalto.fi/projects/rrl/

Sampo-portaalit (14.6.2021). Lisätietoa Sampo-portaaleista ja videoita: https://seco.cs.aalto.fi/applications/sampo/

Semparl-projekti (14.9.2021). Lisätietoa Semanttinen parlamentti -projektista: https://seco.cs.aalto.fi/projects/semparl/

Severi-projekti (13.9.2021). Projektin 2016–2018 kotisivut: https://seco.cs.aalto.fi/projects/severi/

SKT (14.6.2021). J. V. Snellman kootut teokset -verkkopalvelu: http://snellman.kootutteokset.fi/

Sotasampo (14.6.2021). Sotasampo-projektin kotisivu: https://seco.cs.aalto.fi/projects/sotasampo/

Sotasurmat (14..6.2021). Sotasurmat 1914–1922 -projektin kotisivu: https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/

Staab, S., & Studer, R. (2010). *Handbook on Ontologies* (2nd edition). Springer.

SBL (14.6.2021). Svenskt Biografiskt Lexikon -palvelu osoitteessa: https://sok.riksarkivet.se/Sbl/Start.aspx?lang=en

Tamper, M., Leskinen, P., Apajalahti, K., & Hyvönen, E. (2018). Using Biographical Texts as Linked Data for Prosopographical Research and Applications. Teoksessa M. Ioannides et al. (eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection* (pp. 125–137). EuroMed 2018. Lecture Notes in Computer Science, vol 11196. Springer. https://doi.org/10.1007/978-3-030-01762-0_110

Tamper, M., Leskinen, P:, Hyvönen, E., Valjus, R., & Keravuori, K. (2021). Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland. *Semantic Web*, forth-coming. Pre-print: https://seco.cs.aalto.fi/publications/2021/tamper-et-al-bs-2021.pdf

Tunkelang, D. (2009). *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan-Claypool. https://doi.org/10.2200/S00190ED1V01Y200904ICR005

Tuominen, J., Hyvönen, E. & Leskinen, P. (2018). Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. Teoksessa A. Fokkens et al. (eds.), *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)* (pp. 59–66). CEUR Workshop Proceedings, Vol. 2119. http://ceur-ws.org/Vol-2119/

Tzitzikas, Y., Manolis, N. & Papadakos, P. (2017). Faceted exploration of RDF/S datasets: a survey. *Journal of Intelligent Information Systems, 48*(2), 329–364. https://doi.org/10.1007/s10844-016-0413-8

ULAN (14.6.2021). Union List of Artist Names Online: https://www.getty.edu/research/tools/vocabularies/ulan/

Université du Luxembourg (13.9.2021). Digital source criticism: https://ranke2.uni.lu/define-dsc/

Verboven, K., Carlier, M. & Dumolyn, J. (2007). A Short Manual to the Art of Prosopography. Teoksessa K. S. B. Keats-Rohan (ed.), *Prosopography Approaches and Applications. A Handbook* (pp. 35–70). Oxford, Unit for Prosopographical Research (Linacre College).

W3C (14.6.2021). Semanttisen webin standardit, W3C: https://www.w3.org/standards/semanticweb/

Warren, C. (2018). Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics, 3*(1). https://doi.org/10.22148/16.028

Warren, C., Shore, D., Otis, J., Wang, L, Finegold, M., & Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *Digital Humanities Quarterly, 10*(3). http://www.digitalhumanities.org/dhq/vol/10/3/000244/000244.html